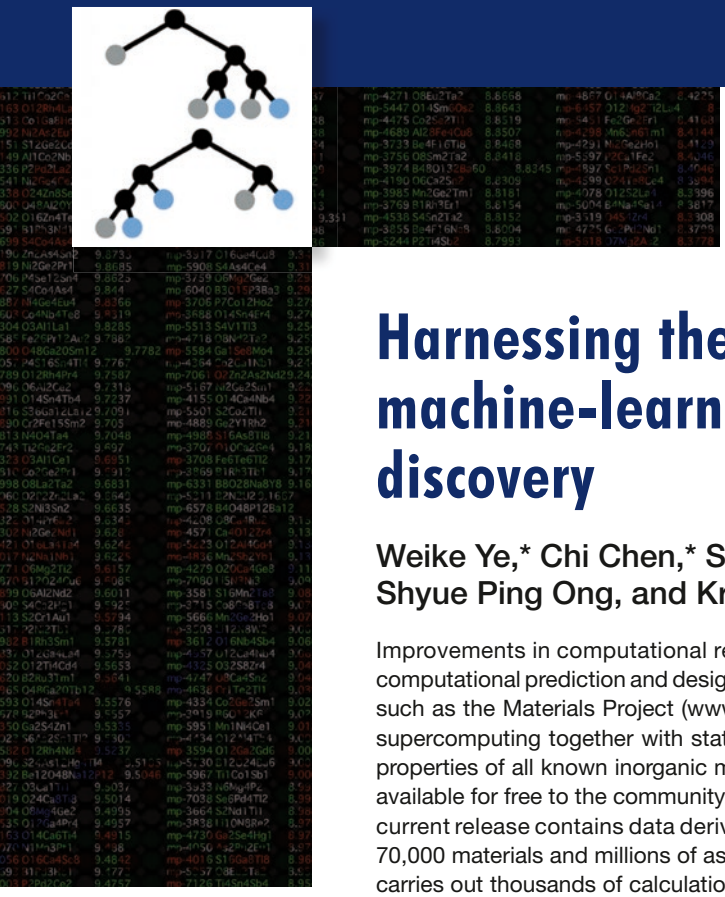


Harnessing the Materials Project for machine-learning and accelerated discovery

Weike Ye,* Chi Chen,* Shyam Dwaraknath,* Anubhav Jain, Shyue Ping Ong, and Kristin A. Persson

Improvements in computational resources over the last decade are enabling a new era of computational prediction and design of novel materials. The resulting resources are databases such as the Materials Project (www.materialsproject.org), which is harnessing the power of supercomputing together with state-of-the-art quantum mechanical theory to compute the properties of all known inorganic materials, to design novel materials, and to make the data available for free to the community, together with online analysis and design algorithms. The current release contains data derived from quantum mechanical calculations for more than 70,000 materials and millions of associated materials properties. The software infrastructure carries out thousands of calculations per week, enabling screening and predictions for both novel solids as well as molecular species with targeted properties. As the rapid growth of accessible computed materials properties continues, the next frontier is harnessing that information for automated learning and accelerated discovery. In this article, we highlight some of the emerging and exciting efforts, and successes, as well as current challenges using descriptor-based and machine-learning methods for data-accelerated materials design.

Introduction

Accelerated data-driven learning aims to reveal hidden relationships in data sets, usually with the end goal of building predictive models that outperform traditional ones. Recently, machine-learning (ML) models have enjoyed success in solving long-standing computer science problems that have eluded direct solution. Examples of such breakthroughs[1] include those in image recognition, speech recognition, and even "scientific" tasks such as analyzing particle accelerator data. The application of ML involves three steps: (1) curation of an input data set of sufficient size and quality, (2) mapping the input data to features/descriptors and targets, and (3) model fitting.

Several features make the Materials Project[2] database uniquely suited for building ML models. First, the variety and quality of data available, both in terms of the properties (e.g., formation energies, energies above the convex hull, bandgaps, elastic constants, and surface energies) as well as the chemistry coverage (most known, unique, crystalline systems in the International Crystallographic Structure Database[3] [ICSD]), are obtained in a self-consistent manner to yield directly comparable properties across structure and chemistry. (The convex hull is the compound energy that prevents decomposition into other compounds or elements in that composition space.) Second, the Materials Project takes great efforts to reduce duplication of data by ensuring the uniqueness of materials and computed properties, which enable fitted ML models to explore structure–property relations without being skewed toward specific, heavily studied structure spaces. This is becoming an increasingly important consideration as the Materials Project and similar databases[4,5] now contain computational data sets ranging on the order of 10^3 materials[6–8] (for computationally expensive property evaluations) to approximately 10^4–10^5 materials[9,10] (for computationally inexpensive property evaluations). Finally, the Materials Project provides a robust Application Programming Interface[11] as well as a

Weike Ye, University of California, San Diego, USA; w6ye@ucsd.edu
Chi Chen, University of California, San Diego, USA; chc273@eng.ucsd.edu
Shyam Dwaraknath, Lawrence Berkeley National Laboratory, USA; shyamd@lbl.gov
Anubhav Jain, Lawrence Berkeley National Laboratory, USA; ajain@lbl.gov
Shyue Ping Ong, University of California, San Diego, USA; ongsp@eng.ucsd.edu
Kristin A. Persson, University of California, Berkeley, and Lawrence Berkeley National Laboratory, USA; kapersson@lbl.gov
*denotes equal contribution.
doi:10.1557/mrs.2018.202

high-level Python (programming language) wrapper in pymatgen,[12] which facilitates the querying of large data sets for ML. Despite the recent and radical increase in available and systematic materials data, there are still challenges in the application of ML models. The most prominent of these are discussed in the following sections.

Data limitations and the use of features/descriptors

One of the historic limitations to applying ML to the materials science domain has been obtaining data sets large and diverse enough to robustly train and validate ML models. Even the Materials Project database is not large enough to use "big data" approaches that include tens of millions of training data points. One way to improve the performance of ML for smaller data sets typically encountered in materials science is to extract features (or descriptors) from the input data. In this context, a set of features (or descriptors) refers to data derived from the original raw inputs that encompass domain knowledge and is expected to simplify the ML problem (**Figure 1**). Depending on the complexity of the problem and the scarcity of available data, the choice of appropriate descriptors can have a large impact on final performance of a ML model.[13]

Early efforts at feature extraction in the materials science domain were concentrated on the chemical composition. The raw string format of a composition (e.g., "$LiFePO_4$") is not amenable to efficient learning because machine-learning algorithms are unaware that this text represents a chemical composition with specific physical characteristics. Studies have instead employed numerical descriptors that represent physical aspects of the composition such as sums of covalent radii, differences in electronegativity, or the average melting point of the component elements in the composition.[14,15] This descriptor generation has been a crucial preprocessing step to machine learning for a wide array of properties, including melting points,[16] thermoelectric figures of merit,[17] thermal conductivity,[9] solute diffusion barriers in face-centered-cubic metals,[18] and elastic properties,[19] to name a few.

One shortcoming of compositional descriptors is the inability to distinguish between entries sharing a chemical formula but possessing different crystal structures (e.g., diamond from graphite or carbon nanotubes). Compositional descriptors alone have had the most success in problems where the crystal structure has been held constant or was the intended variable to predict.[20,21] One active research topic is how to best extract descriptors from a crystal structure which should be invariant to symmetry operations of the crystal. For example, translating all the atoms, rotating the entire crystal, changing the "order" in which atoms are listed, and creating a supercell are all operations that should not change the value of a well-formed crystal structure descriptor. Recently, researchers have devised many such formulations,[22–29] including several variants of a descriptor for the complex network of coulomb interactions within a structure. Faber et al. recently tested this using a data set of 3938 formation energies retrieved from the Materials Project database.[28]

It is also possible to generate features by mixing together combinations of the original features according to various equations,[30,31] (e.g., as the sum of two features divided by a third). Although it is possible to generate many thousands of features or more, selecting the most relevant compound features is still an area of active research. The presence of redundant or uninformative features can reduce the overall performance of machine learning, making it prone to overfitting (i.e., reproducing patterns present in the known data that will not accurately reflect new data). Various feature selection methods devised by the ML community[32] and certain ML algorithms such as kernel ridge regression employ regularization parameters that can be used to penalize the use of too many features.

Although most descriptors are constructed manually using domain knowledge or some mathematical basis, it is also possible to automatically construct descriptors with sufficient data. Representation learning[32] tackles the issue of how to best represent raw ML inputs. One approach is an autoencoder, which is a neural network architecture designed to transform the original input into a reduced set (or a latent representation) that can faithfully be transformed back into the full input. This method can be used to generate representations on a large unlabeled data set and transfer this representation model to a smaller but similar data set. For example, Google trained an autoencoder model on 10 million image thumbnails to uncover features that corresponded to "face descriptors" or "cat descriptors" from unlabeled data.[33] These automatically generated features, which did not require domain knowledge to devise, were used to boost the performance of an image

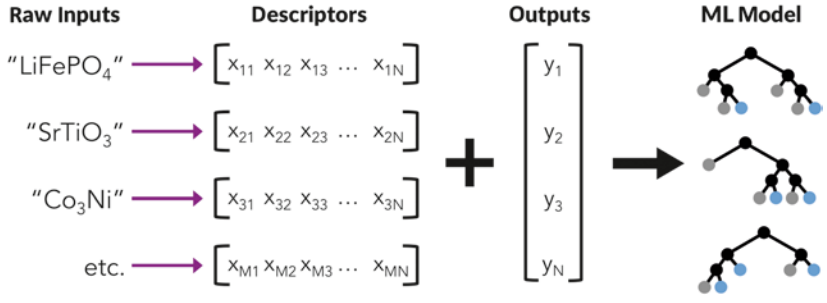


Figure 1. Example of machine-learning (ML) workflow, including the generation of descriptors. Each raw input is transformed into a set of descriptors. Here, all raw inputs are chemical compositions; they might instead be a crystal structure, experimental spectrum, or microstructural image. The descriptors represent a set of numbers that encapsulate the raw inputs in a compact and physically meaningful way. For example, one descriptor of a composition might be the variation in electronegativity of its component elements, which would indicate the degree to which the compound might be metallic, covalent, or ionic. The set of descriptors (sometimes in conjunction with the raw inputs) and known outputs are then used to efficiently train the ML model.

classifier by more than 70% over the existing state-of-the-art models on the well-known Imagenet[34] data set. This type of approach is termed "transfer learning" and variants of this approach have recently been applied to the materials science domain as well.[35] Such research hints that it might be increasingly possible in the future to "auto-learn" descriptors given large and sufficiently diverse data sets.

ML and accelerated discovery

An attractive feature of ML models is their speed, especially compared to obtaining new materials property data using either experiments or computations. In this section, we highlight several examples of applying ML models to computed materials data. As an example, phase stability presents a ubiquitous property of interest in materials science. Faber and co-workers[36] applied kernel ridge regression to predict the formation energies of two million elpasolite (ABC_2D_6) structures with a mean average error of approximately 100 meV/atom. Considering that approximately 90% of the crystals in the ICSD have energy above the hull (E_{hull}) < 70 meV/atom,[37] and the errors of density functional theory (DFT)-calculated formation energies of ternary oxides from binary oxides relative to experiments are ~24 meV/atom,[38] there is still significant room for improvement in fit-to-DFT stability predictions. The choice of the target metric in particular may not be the most appropriate for ML applications. Both, E_{hull} and the formation energy (E_f) (referenced to elemental compounds or binary oxides) are not ideal learning targets due to the variation in reference states for each species across different structures, which introduces additional DFT errors that can bias the learning and discovery process.[39] Faber et al. identified 128 new expected stable elpasolite structures, 90 of which were unique stoichiometries and could not have been identified in the same timeframe using direct DFT computation, despite the lack of accuracy.

Bandgap is another important material property commonly estimated via first-principles calculations. The 54,000 band structures in the Materials Project are calculated by DFT with the generalized gradient approximation (GGA), which underestimates the bandgap due to the approximation in exchange-correlation functionals, the self-interaction error, and the missing derivative discontinuity. Common solutions to this problem rely on more accurate, but more computationally expensive, methods such as the modified Becke–Johnson (mBJ) functional,[40] the delta self-consistent-field (ΔSCF) method,[41] hybrid functionals (HSE06)[42] or GW calculations based on many-body perturbation theory.[43] Lee and co-workers,[44] attempted to circumvent this by learning from the cheaper PBE bandgap (E_g [PBE]) or mBJ bandgap (E_g [mBJ]), with an additional 18 descriptors to predict the more accurate G_0W_0 bandgap (E_g (G_0W_0)). The best model trained by nonlinear support vector regression (SVR) yielded a root-mean-square (rms) error of 0.24 eV (**Figure 2**).

Another important set of materials properties is mechanical behavior. The elastic bulk and shear moduli (K and G, respectively) govern the stress–strain relations of isotropic materials, and are highly correlated to properties such as hardness[45] and thermal conductivity.[46] Recently, 1940 computed elastic constants from the Materials Project were utilized as a training set for a gradient boosting machine local polynomial regression (GBM-Locfit) model to predict K and G for k-nary inorganic polycrystalline compounds by de Jong.[19] After careful optimization of hyper-parameters through tenfold cross-validation, the rms error of K and G were 0.0750 and 0.1378 (log(GPa)), respectively (**Figure 3**). The authors found that for both K and G, the top four most important descriptors were V (volume per atom), R_n (elemental row number), E_c (cohesive energy), and X (mean of elemental electronegativity). The model was therefore used to screen for hard materials among 30,000 compounds from the Materials Project, with $Mg(B_6C)_2$, Sc_2CrB_6, and $Mg_2B_{24}C$ discovered as promising candidates.[19] This work is a demonstration of how data from the Materials Project can be directly used to extract useful information and inspire new high-throughput screening.

Besides predicting basic materials properties, ML can also assist in developing an understanding of the fundamental physics of crystals by constructing force fields. Force-field

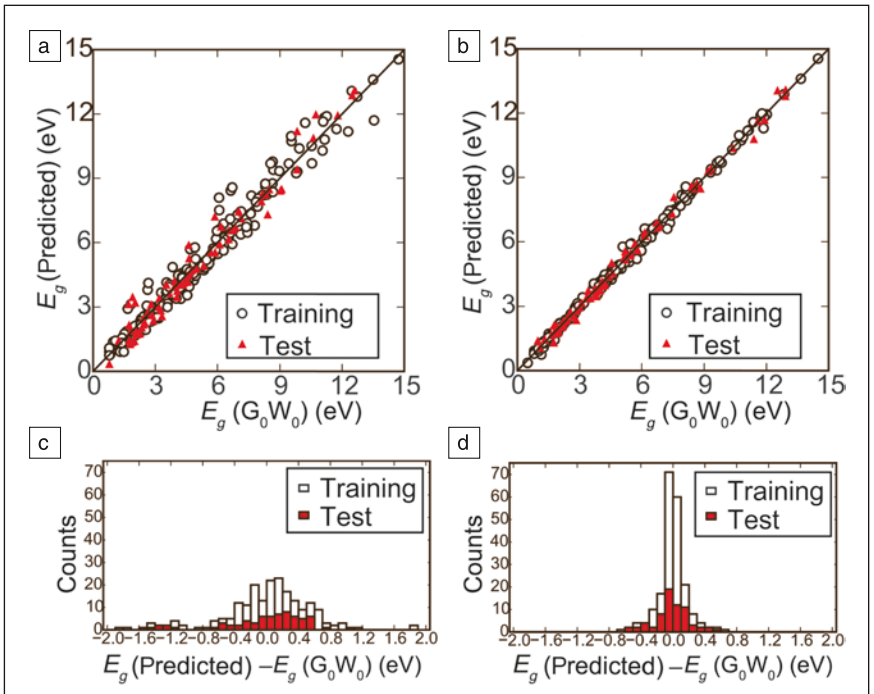


Figure 2. Comparison of the E_g (G_0W_0) method and E_g (predicted). (a, c) The ordinary least-squares regression model with E_g (PBE) as a single predictor. (b, d) The support vector regression model with the 18-predictor set. A randomly chosen result from 100 trials is shown. Reprinted with permission from Reference 44. © 2016 IOP Publishing.

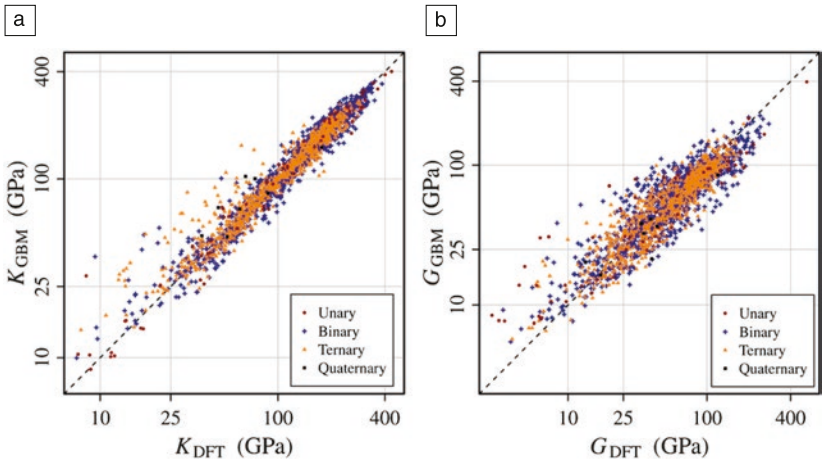


Figure 3. Comparison of density functional theory (DFT) training data with GBM-Locfit predictions for *K* (a) and *G* (b). The training set consists of 65 unary, 1091 binary, 776 ternary, and eight quaternary compounds.[19]

development allows for the simulation of systems that contain thousands of atoms and over time scales of nanoseconds to microseconds. Chen and co-workers[47] applied ML techniques to build a spectral neighbor analysis potential (SNAP) model for Mo, which outperforms existing Mo potentials in accuracy over energies, stress tensors, elastic properties, melting point, and surface and grain-boundary energies using the data and software tools developed by the Materials Project. The data set was further screened by a principal component analysis, which ensures distinct local environments among data. A differential evolution algorithm was employed to simultaneously optimize the weights of the ML model and spectrum parameters used to calculate features. This work, besides presenting a superior force field for Mo, showcases inspiring strategies in data preparation and global optimization.

It is also worth noting that the fruit of ML can in return empower the Materials Project. Simultaneously introduced with XASdb, a database containing ~800,000 *K*-edge XANES spectra, a novel ensemble-learned spectrum identification algorithm was proposed to enable instant matching between experimental XAS spectra and the references in XASdb.[48] The algorithm classifies an input spectrum according to chemistry and oxidation state, using a group of 33 weakly correlated algorithms in an ensemble akin to a random forest, to span the whole space of materials chemistry and oxidation states available in the Materials Project. This divide-and-conquer approach correctly identified the oxidation state and coordination environment with 84.2% accuracy in a test of 19 high-quality experimental spectra and is currently implemented in the Materials Project for rapid classification of measured XAS spectra. The combination of the database and the designed algorithm significantly accelerates materials characterization using XAS and XANES spectra, and represents a powerful addition to the Materials Project.

Finally, a recent development is the application of deep learning to materials science. Xie and coworkers[49] developed convolutional neural networks (CNN) to learn from the "crystal graph," which represents the crystal structure using graph nodes and edges that represent atoms and bonds. A Materials Project data set containing 46,744 materials, covering 87 elements, seven lattice systems, and 216 space groups, with seven properties (DFT calculated formation energy, absolute energy, bandgap, Fermi energy, bulk moduli, shear moduli, and the Poisson's ratio) was used for training. The results show that with careful design of crystal representation, CNN models can be generalized to predict a wide range of properties with a level of accuracy comparable to DFT. The effort of generalizing both the structure and target space was greatly emphasized in this work and should be acknowledged. Another merit in the developed CNN model is the interpretability, meaning insights can be drawn from analyzing the contribution to the target from different local environments of atoms, which were encoded in the input vectors.

Outlook

The discussed examples showcase the power of combining robust and diverse materials data with ML techniques. Besides directly feeding existing data or initial structures for model training, the Materials Project database and its related infrastructures for structural analysis[12] and high-throughput calculations[50] are also enabling efficient generation of new data[51] for specific learning tasks. We emphasize that while increasing accuracy is always a goal of any ML study, efforts should focus on increasing the scale, diversity and quality of data. For example, while the few hundred data points, as used in Lee's work[44] on bandgap prediction, as well as Medasani's work[52] on classifying defect types in B2 intermetallics were used to extract trends, such sets are insufficient for the automatic training of robust ML models. Furthermore, most models are only applicable within constrained structure space,[36,53,54] but a truly robust model across most structure types is still awaited. Hence, until enough consistent and reliable materials data is available across structure and chemical spaces, the key may lie in the development of better compact yet descriptive structure descriptors.

The forward approach adopted by the Materials Project is to support both avenues, by (1) continuously generating and making available ever-growing data sets of fundamental materials properties across diverse chemistries and structures, and (2) developing algorithms to generate descriptor functions that can be used in ML methods for smaller data sets. The Materials Project has been continually releasing new types of large data sets that can serve as the basis for future machine learning studies.[7,8,55,56] Further, the Materials Project is collaborating on the development of the matminer (www.github.com/hackingmaterials/matminer) software

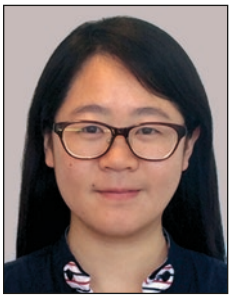
platform, which provides open-source implementations in the Python programming language for a variety of descriptors of all types that have been proposed in the literature. For example, the Materials Project has recently added order parameter functions that are highly sensitive to the local environment (e.g., a "tetrahedral" versus "square planar" arrangement of atoms around a central site[57]), and can be used to assess structure similarity or improve structure prediction performance. Our hope is that an abundant user base, a large well-curated database of properties and descriptors, combined with a community that is enthusiastic to employ machine learning in materials science will foster the next generation of advances in structure–property relations and materials discovery at an ever-increasing pace.

Acknowledgments

S.D., A.J., and K.P. acknowledge support by the US Department of Energy (DOE) Basic Energy Sciences (BES) Materials Project Program, Contract No. KC23MP. W.Y., C.C., and S.P.O. acknowledge support from the Samsung Advanced Institute of Technology's (SAIT) Global Research Outreach (GRO) Program for this work.

References

1. Y. Lecun, Y. Bengio, G. Hinton, *Nature* **521**, 436 (2015).
2. A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, *APL Mater.* **1**, 011002 (2013).
3. G. Bergerhoff, R. Hundt, R. Sievers, I.D. Brown, *J. Chem. Inf. Comput. Sci.* **23**, 66 (1983).
4. S. Kirklin, J.E. Saal, B. Meredig, A. Thompson, J.W. Doak, M. Aykol, S. Ruhl, C. Wolverton, *NPJ Comput. Mater.* **1**, 15010 (2015).
5. S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R.H. Taylor, L.J. Nelson, G.L.W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, O. Levy, *Comput. Mater. Sci.* **58**, 227 (2012).
6. A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, I. Tanaka, *Phys. Rev. Lett.* **115**, 205901 (2015).
7. M. de Jong, W. Chen, H. Geerlings, M. Asta, K.A. Persson, *Sci. Data* **2**, 150053 (2015).
8. M. de Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C.K. Ande, S. van Der Zwaag, J.J. Plata, C. Toher, S. Curtarolo, G. Ceder, K.A. Persson, M. Asta, *Sci. Data* **2**, 150009 (2015).
9. J. Carrete, W. Li, N. Mingo, S.D. Wang, S. Curtarolo, *Phys. Rev. X* **4**, 011019 (2014).
10. F. Ricci, W. Chen, U. Aydemir, G.J. Snyder, G.-M. Rignanese, A. Jain, G. Hautier, *Sci. Data* **4**, 170085 (2017).
11. S.P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, K.A. Persson, *Comput. Mater. Sci.* **97**, 209 (2015).
12. S.P. Ong, W.D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V.L. Chevrier, K.A. Persson, G. Ceder, *Comput. Mater. Sci.* **68**, 314 (2013).
13. S.S. Young, F. Yuan, M. Zhu, *Mol. Inform.* **31**, 707 (2012).
14. A. Seko, A. Togo, I. Tanaka, *Nanoinformatics* **3**, 23 (2018).
15. L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, *NPJ Comput. Mater.* **2**, 16028 (2016).
16. A. Seko, T. Maekawa, K. Tsuda, I. Tanaka, *Phys. Rev. B Condens. Matter Mater. Phys.* **89**, 1 (2014).
17. J. Carrete, N. Mingo, S. Wang, S. Curtarolo, *Adv. Funct. Mater.* **24**, 7427 (2014).
18. H. Wu, A. Lorenson, B. Anderson, L. Witteman, H. Wu, B. Meredig, D. Morgan, *Comput. Mater. Sci.* **134**, 160 (2017).
19. M. de Jong, W. Chen, R. Notestine, K. Persson, G. Ceder, A. Jain, M. Asta, A. Gamst, *Sci. Rep.* **6**, 34256 (2016).
20. F. Legrain, J. Carrete, A. van Roekeghem, S. Curtarolo, N. Mingo, *Chem. Mater.* **29**, 6220 (2017).
21. M.W. Gaultois, A.O. Oliynyk, A. Mar, T.D. Sparks, G.J. Mulholland, B. Meredig, *APL Mater.* **4** (5), 05312 (2016).
22. O. Isayev, C. Oses, S. Curtarolo, A. Tropsha, *Nat. Commun.* **8**, 15679 (2017).
23. O. Isayev, D. Fourches, E.N. Muratov, C. Oses, K. Rasch, A. Tropsha, S. Curtarolo, *Chem. Mater.* **27**, 735 (2015).
24. A.P. Bartók, M.C. Payne, R. Kondor, G. Csányi, *Phys. Rev. Lett.* **104**, 1 (2010).
25. A.P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B Condens. Matter Mater. Phys.* **87**, 1 (2013).
26. H. Huo, M. Rupp, (2017), https://www.researchgate.net/profile/Matthias_Rupp/publication/316429137_Unified_Representation_for_Machine_Learning_of_Molecules_and_Crystals/links/597ba6ff0f7e9b880293f6bf/Unified-Representation-for-Machine-Learning-of-Molecules-and-Crystals.pdf.
27. J. Behler, M. Parrinello, *Phys. Rev. Lett.* **98**, 146401 (2007).
28. F. Faber, A. Lindmaa, O.A. von Lilienfeld, R. Armiento, *Int. J. Quantum Chem.* **1** (115), 1094 (2015).
29. T.L. Pham, H. Kino, K. Terakura, T. Miyake, I. Takigawa, K. Tsuda, H.C. Dam, *Sci. Technol. Adv. Mater.* **18** (1), 756 (2017).
30. L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, M. Scheffler, *Phys. Rev. Lett.* **114** (10), 105503 (2015).
31. R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, *NPJ Comput. Mater.* **3** (1), 54 (2017).
32. G. Chandrashekar, F. Sahin, *Comput. Electr. Eng.* **40**, 16 (2014).
33. Q.V. Le, *Proc. 2013 IEEE Int. Conf. Acoust. Speech Signal Process.* (2013), p. 8595.
34. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, *2009 IEEE Conf. Comput. Vis. Pattern Recognit.* (2009), p. 248.
35. M.L. Hutchinson, E. Antono, B.M. Gibbons, S. Paradiso, J. Ling, B. Meredig, ArXiv preprint arXiv:1711.05099 (2017).
36. F.A. Faber, A. Lindmaa, O.A. Von Lilienfeld, R. Armiento, *Phys. Rev. Lett.* **117**, 135502 (2016).
37. W. Sun, S.T. Dacek, S.P. Ong, G. Hautier, A. Jain, W.D. Richards, A.C. Gamst, K.A. Persson, G. Ceder, *Sci. Adv.* **2**, e1600225 (2016).
38. G. Hautier, S.P. Ong, A. Jain, C.J. Moore, G. Ceder, *Phys. Rev. B Condens. Matter Mater. Phys.* **85**, 155208 (2012).
39. L. Wang, T. Maxisch, G. Ceder, *Phys. Rev. B Condens. Matter* **73**, 195107 (2006).
40. F. Tran, P. Blaha, *Phys. Rev. Lett.* **102**, 226401 (2009).
41. M.K.Y. Chan, G. Ceder, *Phys. Rev. Lett.* **105**, 196403 (2010).
42. J. Heyd, G.E. Scuseria, M. Ernzerhof, *J. Chem. Phys.* **118**, 8207 (2003).
43. F. Fuchs, J. Furthmüller, F. Bechstedt, M. Shishkin, G. Kresse, *Phys. Rev. B Condens. Matter* **76**, 115109 (2007).
44. J. Lee, A. Seko, K. Shitara, K. Nakayama, I. Tanaka, *Phys. Rev. B Condens. Matter* **93**, 115104 (2016).
45. S.F. Pugh, *London Edinburgh Dublin Philos. Mag. J. Sci.* **45**, 823 (1954).
46. G.J. Snyder, E.S. Toberer, *Nat. Mater.* **7**, 105 (2008).
47. C. Chen, Z. Deng, R. Tran, H. Tang, I.-H. Chu, S.P. Ong, *Phys. Rev. Mater.* **1**, 043603 (2017).
48. C. Zheng, K. Mathew, C. Chen, Y. Chen, H. Tang, A. Dozier, J.J. Kas, F.D. Vila, J.J. Rehr, L.F.J. Piper, K. Persson, S.P. Ong, *NPJ Comput. Mater.* **4** (1), 12 (2018).
49. T. Xie, J.C. Grossman, *Phys. Rev. Lett.* **120**, 145301 (2018).
50. A. Jain, S.P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier, D. Gunter, K.A. Persson, *Concurr. Comput.* **27**, 5037 (2015).
51. K. Mathew, J.H. Montoya, A. Faghaninia, S. Dwaraknath, M. Aykol, H. Tang, I.-H. Chu, T. Smidt, B. Bocklund, M. Horton, J. Dagdelen, B. Wood, Z.K. Liu, J. Neaton, S.P. Ong, K. Persson, A. Jain, *Comput. Mater. Sci.* **139**, 140 (2017).
52. B. Medasani, A. Gamst, H. Ding, W. Chen, K.A. Persson, M. Asta, A. Canning, M. Haranczyk, *NPJ Comput. Mater.* **2** (1), 1 (2016).
53. J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, M.A.L. Marques, *Chem. Mater.* **29**, 5090 (2017).
54. G. Pilania, A. Mannodi-Kanakkithodi, B.P. Uberuaga, R. Ramprasad, J.E. Gubernatis, T. Lookman, *Sci. Rep.* **6**, 19375 (2016).
55. I. Petousis, D. Mrdjenovich, E. Ballouz, M. Liu, D. Winston, W. Chen, T. Graf, T.D. Schladt, K.A. Persson, F.B. Prinz, *Sci. Data* **4**, 160134 (2017).
56. G. Petretto, S. Dwaraknath, H.P.C. Miranda, D. Winston, M. Giantomassi, M.J. Van Setten, X. Gonze, K.A. Persson, G. Hautier, G.-M. Rignanese, *Sci. Data* **5**, 180065 (2018).
57. N.E.R. Zimmermann, M.K. Horton, A. Jain, M. Haranczyk, *Front. Mater.* **4**, 1 (2017). □



Weike Ye is a doctoral candidate in the Materials Virtual Lab at the University of California, San Diego. She received her master's degree in 2016 from the Department of Chemistry and Biochemistry at the University of California, San Diego, and her bachelor's degree in 2014 from Nanjing University, China. Her research focuses on applying data analyzing techniques to reveal fundamental structure–property relationships in solid-state materials. Ye can be reached by email at w6ye@ucsd.edu.

Chi Chen is a postdoctoral associate at the University of California, San Diego. He obtained his bachelor's degree from the University of Science and Technology of China. He received his PhD degree from The Hong Kong University of Science and Technology, working on materials simulations and method developments. His current research focuses on applying machine learning to materials science. Chen can be reached by email at chc273@eng.ucsd.edu.

Shyam Dwaraknath is a materials research scientist at Lawrence Berkeley National Laboratory. His research focuses on developing better metrics for synthesizability and improving synthesis outcomes using a combination of theory and machine learning. He received his BS degree in nuclear engineering as well as electrical engineering and computer science from the University of California, Berkeley, and his PhD degree from the University of Michigan. Dwaraknath can be reached by email at shyamd@lbl.gov.

Anubhav Jain is a research scientist/chemist focusing on new materials discovery using high-throughput computations at the Lawrence Berkeley National Laboratory. He received his BE degree in applied and engineering physics from Cornell University and his PhD degree from the Massachusetts Institute of Technology. He is a recipient of the DOE Computational Science Graduate Fellowship, the Luis W. Alvarez Postdoctoral Fellowship, and a DOE Early Career Award. Jain can be reached by email at ajain@lbl.gov.

Shyue Ping Ong is an associate professor at the University of California, San Diego. He received his PhD degree in materials science and engineering from the Massachusetts Institute of Technology in 2011. While completing his PhD degree, he was part of a research team that developed a high-throughput computational framework for materials design that led to the discovery of several novel Li-ion battery cathode materials. In 2011, this framework was spun off into the Materials Project, an open science project to make the calculated properties for all known inorganic materials publicly available to researchers to accelerate materials innovation. Ong can be reached by email at ongsp@eng.ucsd.edu.

Kristin Persson is a tenured associate professor of materials science and engineering at the University of California, Berkeley, with a joint appointment as faculty staff scientist at the Lawrence Berkeley National Laboratory, where she leads several research teams, including the Materials Project. She uses computational methods and machine learning to advance materials for clean energy. She co-founded the clean energy startup Pellion Technologies Inc., received numerous awards, including the TMS Early Career Faculty Fellow Award, holds several patents, and has co-authored more than 100 publications. Persson can be reached by email at kapersson@lbl.gov.